# Finding the Data: Enhancing the Utility of the LTER Controlled Vocabulary

1. *Vision and Summary:* **Scientists seeking data should be able to efficiently and reliably locate LTER datasets through searching, browsing or following links from non-LTER systems.** Unfortunately, efficient and reliable searches are not currently possible because of eclectic application of keywords in LTER metadata. Most existing keywords are used by only a single LTER site, or worse yet, in only a single dataset, making them unsuitable for identifying data for cross-site syntheses. To address this problem, the LTER Information Management Committee (IMC) created a working group to help create a controlled vocabulary for use in LTER datasets (<http://databits.lternet.edu/spring-2010/controlled-vocabulary-lter-datasets>). Following recommendations in ANSI/NISO Z39.19 (Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies), the group created a list of 640 preferred terms and associated synonyms that maximized overlap with external vocabularies (e.g., NBII Thesaurus) and application across LTER sites. In early 2010 the LTER Executive Board committed to helping to locate some domain scientists to work with the Information Management Committee on future activities related to keywords, and endorsed the use of the list by LTER sites. Subsequently, the LNO, in association with the “Helping Interdisciplinary Vocabulary Engineering” (HIVE) project, developed some web-services and a prototype client that scans EML metadata and suggests appropriate keywords.

During the 2010 IMC meeting, the VOCAB working group identified five major actions that were needed to realize the vision of efficient and reliable identification of LTER datasets. These actions were:

* 1. **Adopt organizational structures and protocols for managing improvements to the LTER Controlled Vocabulary**
  2. **Develop a database and associated tools for managing the controlled vocabulary**
  3. **Create taxonomys (hierarchies) for keywords that can be used to enhance browsing**
  4. Create tools to aid in adding keywords from the controlled vocabulary to existing and future metadata, and
  5. Develop search and browse tools that exploit the controlled vocabulary

The VOCAB product-oriented working group will focus on actions “a” through “c.” Actions “d” and “e” are dependent upon the other actions and so can be deferred until more progress has been made.

1. *Scope of work:* Products of this working group will be applicable to all LTER sites and will enhance the overall functionality of the NIS (you can’t integrate data you can’t find). They will also be helpful in enhancing existing search and browse capabilities for data systems.
2. *Products:* Products of the proposed activities will include:
   1. **Terms of Reference** that define how the controlled vocabulary and its derivative products will be managed.
   2. **A database of keywords** that implements a data model based on “11.1.4 Term Records” in NISO/ANSI Z39.19. The database will accommodate addition of relationships needed to implement taxonomys (hierarchies) of keywords and synonym rings. Additionally, it will allow a “scope” to be applied to each term, so that individual site, as well as LTER-wide keywords can be accommodated. This database will be a critical resource for keywording and search tools, as well as facilitating management of the controlled vocabulary and creation of taxonomys of relationships.
   3. **Multiple taxonomys (hierarchies) will be developed** that place keywords into a broader context. For example, LTER core areas can be used as high-level concepts in a “research topic” taxonomy. Other taxonomys might address “biomes,” “biota” or “biogeochemistry.” In addition to fully implementing several taxonomys, we will also identify other potential taxonmys that might be created in the future.
3. *Tasks:* **Intensive Development Workshop.**  The goals of the workshop , to be held in February 2011, will be: 1) to define use cases for the keyword database and its development, 2) to identify suitable existing database structures or software for managing the controlled vocabulary and adopt or modify them to meet the use cases, and 3) develop procedures for capturing and managing keyword taxonomys. The workshop will consist of 4 “visiting” participants, 1 “local” participant. During the workshop, time will be scheduled for VTCs to communicate ideas and results and get feedback from “virtual” participants. The workshop will be 3-days in length with travel days before and after. Preparation for the workshop will include formulating a draft terms of reference document that defines the institutional structures and responsibilities for keyword-related activities, that will be reviewed during the workshop and refined based on the use cases. Prior to the workshop candidate data models or software will be identified for further evaluation during the workshop.

**Development of the taxonomys** will be conducted in coordination with the May 2011 Science Council (SC) Meeting. Prior to the workshop, LTER IM’s will create several “straw man” taxonomys using the tools and database identified in the first workshop. An “in person” workshop involving three IM’s and two domain scientists attending the SC meeting (who will be identified by the Executive Board, per the prior agreement) will be held on the day prior to the start of the SC meeting. The IM’s will continue work on the polytaxonomys, incorporating feedback from the designated domain scientists and other SC meeting attendees during the SC meeting, with the goal of having completed several taxonomys by the end of the SC meeting. The resulting taxonomys will be both immediately useful in helping to organize data for use with the NIS, and have long-term utility in helping to refine the processes needed to create additional taxonomys.

5. *Participants*: Participants will be drawn from membership of the VOCAB group. VOCAB chair John Porter will participate in both workshops to help provide continuity. Duane Costa will be a valuable participant in the intensive workshop, which links to his HIVE work. Either as an in-person or virtual participant as will Kristen Vanderbilt, who volunteered during the IMC meeting, and Inigo San Gil from the LNO has also expressed an interest. Additional IMs with expertise in data modeling and/or web services will be selected. IMC co-chairs Margaret O’Brien and Don Henshaw have volunteered to participate in the taxonomy development activity held in association with the SC meeting, and several other IMs have expressed an interest in this activity.

1. *Budget*:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Activity** | **Number of "in person" Participants** | **Airfares @$600** | **Travel Cost** | **Food & Lodging Person Days** | **Food & Lodging Cost @ $77+$46** | **Meeting Costs** | **Total Cost** |
| Database Development | 5 | 4 | $2,400 | 15 | $1845 | $100 | $4,345 |
| Polytaxonomy Meeting | 5 | 2 | $1,200 | 10 | $1230 | $100 | $2,530 |
| **Totals** | **10** | **6** | **$3,600** | **25** | **$3,075** | **$200** | **$6,875** |

1. *Budget Justification*: Airfares were estimated to average $600 per attendee. Lodging and food costs were estimated at $77 lodging + $46 food per diem. $100 for incidental meeting costs (e.g,. the traditional caffine and twinkies needed to keep ecoinformaticians operating, and meeting supplies) were requested. For the Polytaxonomy meeting, to be held in association with the SC meeting, domain scientists are assumed to be traveling to the meeting using other travel funds, but will need an extra day of lodging and food.

Submitted by John Porter on behalf of the VOCAB working group.